Hotel Booking Analytics

Authors:

- Josiah Lee En (National University of Singapore)
- Lee Jin (National University of Singapore)
- Chermane Goh (National University of Singapore)
- Kevin Christian (National University of Singapore)
- Teo Ming Jun (National University of Singapore)

Setup

```
# install.packages("ggrepel")
library(tidyverse)
library(lubridate)
library(ggrepel)
library(ggplot2)
set.seed(43)

dataset_link <- paste0("https://raw.githubusercontent.com/rfordatascience/",
"tidytuesday/master/data/2020/2020-02-11/hotels.csv")

hotels <- readr::read_csv(dataset_link)
```

Introduction

Across many studies, Data Analytics has shown to be key to an organization's success(McAfee A, et al. 2012). With this in mind, our group would like to harness data from the Hotel bookings demand dataset(N. Antonio, A. De Almeida, and L. Nunes, 2019) to help those in the hotelier industry make more informed business decisions. The first question we have in mind is how does room availability, time of year and country of guests affect the number of guests. We believe this question would allow readers to find out more about the factors that affect the number of guests staying in the hotels. The next question is what are the factors which affect cancellation rates in hotel bookings. This finding will allow readers to understand more about the causes behind hotel booking cancellations. Both of these findings are significant for hotel business supervisors when it comes to business operations. Specifically, this data would be greatly beneficial to hoteliers to help them plan more resources for a peak season as well as to reduce cancellation rates.

Data Description

What the data is about:
The hotel booking demand dataset contains bookings due to arrive between the 1st of July 2015 and the 31st of August 2017, focusing on two hotels in Portugal, a resort hotel in Algarve(H1) and a city hotel in Lisbon(H2). Each row corresponds to the booking of one room.

Transformation/Cleaning:
Our group noticed there were some entries where the number of adults was extremely large. Upon investigation, all bookings with more than 4 adults had been cancelled. Since each booking row corresponds to one room as given by the variable room_type, we conjectured that these bookings had been mistakenly keyed in for group bookings. We thus filtered out bookings with more than 4 adults. Next, we convert the 3 columns arrival_date_year, arrival_date_month, arrival_date_day_of_month into a date type column via lubridate to allow us to explore the dataset easily.

```
# Max number of adults for non-cancelled bookings
hotels %>%
  filter(is_canceled==0) %>%
  summarise(max(adults))
```

```
## # A tibble: 1 x 1
##   `max(adults)`
##           <dbl>
## 1             4
```

```
# Convert the following 3 columns into arrival_date:
# - arrival_date_year,
# - arrival_date_month
# - arrival_date_day_of_month
df=hotels %>%
  filter(adults<=4) %>%
  mutate(
    arrival_date_str=paste(
      arrival_date_year,
      arrival_date_month,
      arrival_date_day_of_month
    )
  ) %>%
  mutate(arrival_date=ymd(arrival_date_str)) %>%
  select(-c(arrival_date_year,arrival_date_month,arrival_date_day_of_month))
```

Key descriptive statistics: Diving into the data, the first thing that our group wanted to find out was how many bookings were made for each hotel. The code is as follows.

```
# City Hotel has about twice the bookings of Resort Hotel
df %>%
  group_by(hotel) %>%
  summarise(n=n())
```

```
## # A tibble: 2 x 2
##   hotel            n
##   <chr>        <int>
## 1 City Hotel   79330
## 2 Resort Hotel 40044
```

Question 1: How does room availability, time of year and country of guests affect number of arrivals?

Introduction

Being able to capture the trends allows hotels to better plan their logistics and staff, providing an overall enhanced experience for their valued guests. We picked out 3 factors that we think might affect the number of guest arrivals. We decided to focus on non-cancelled bookings in the year 2016. This is so that we can see the trend over 12 months and because data for the years 2015 and 2017 are incomplete.

Methodology

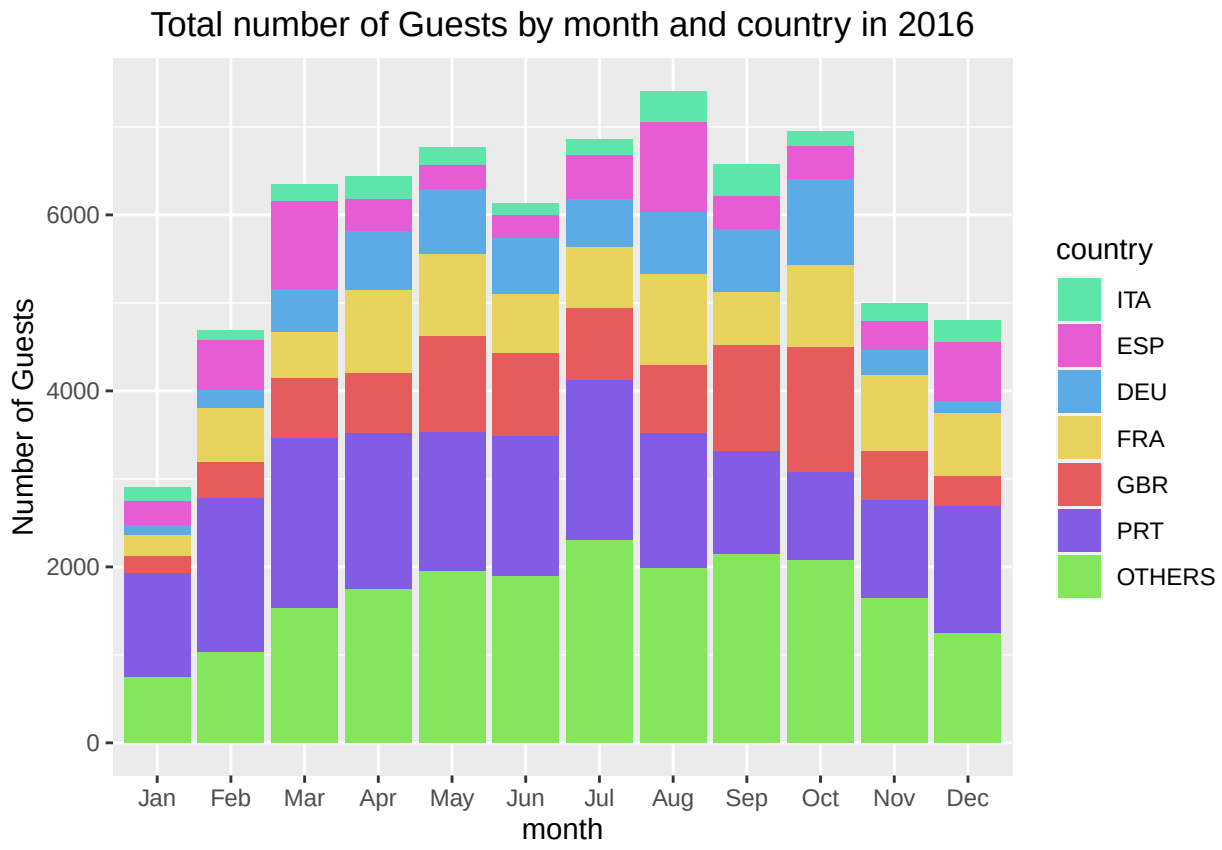
Plot1 We filter out non-cancelled reservations for the year 2016. Next, we picked out the top 6 countries with the most guests. We then grouped all other countries under "OTHERS". Finally, we calculated the total number of guests by month and by country. For our plot, we used a barplot of the number of arrived guests vs month. We split this by country using fill. We also order this split by the number of guests ascending so we have countries with most guests at the bottom, least on top. This makes it easier for the reader to visualise changes.

Plot2 We first calculate nights_stayed, which is the sum of stays_in_week_nights and stays_in_weekend_nights. We note that some bookings have 0 nights stayed. We decided to filter them out as hotels likely use day only bookings as a way to fill up empty rooms. Then, we determine the start and end date of each reservation. Finally, we count the number of rooms that are occupied every day of the year. We use a line plot of the number of rooms occupied vs days, split by hotel using colour. We also changed the breaks of the x-axis to match the middle of each month. Finally, we added 2 dashed lines to indicate the maximum number of rooms occupied for both hotels.

Visualizations

```
# Plot 1
df1 = df %>%
  filter(is_canceled==0) %>%
  filter(year(arrival_date)==2016) %>%
  mutate(guests = adults+children+babies)
temp = df1 %>%
  group_by(country) %>%
  summarize(n=sum(guests))
top6 = temp %>%
  slice_max(order_by=n, n=6)
top6countries = top6$country
temp2 = df1 %>%
  filter(country %in% top6countries)
temp3 = df1 %>%
  filter(!(country %in% top6countries)) %>%
  mutate(country = "OTHERS")
df2 = bind_rows(temp2, temp3) %>%
  mutate(month=month(arrival_date, label=TRUE)) %>%
  group_by(month, country) %>%
  summarise(n=sum(guests))
cols = rainbow(7, s=.6, v=.9)[sample(1:7,7)]
ggplot(df2, aes(x=month, y=n, fill=reorder(country, n))) +
```

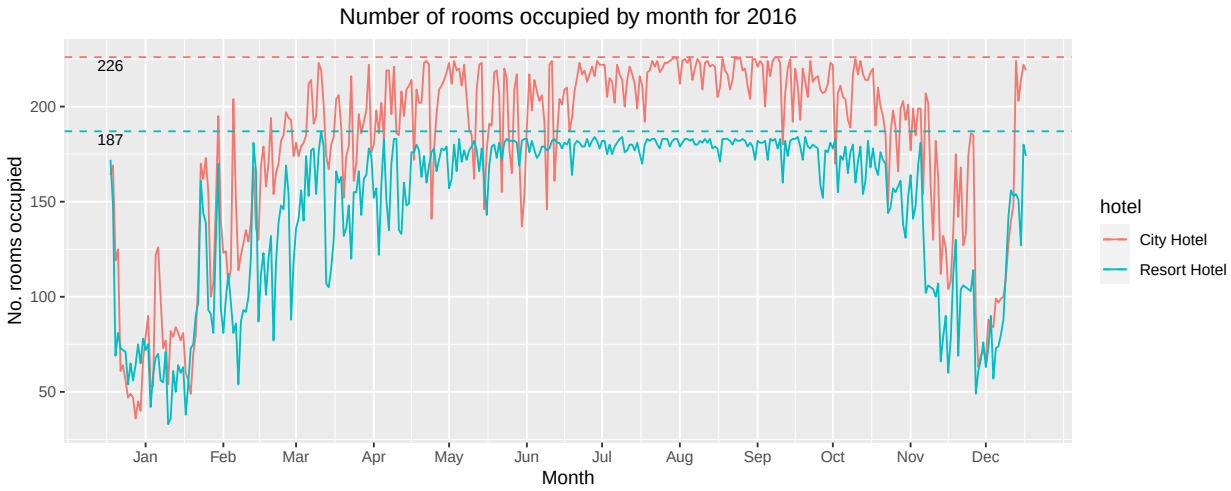
```
  geom_col() +
  scale_fill_manual(values=cols) +
  labs(
    y="Number of Guests",
    title="Total number of Guests by month and country in 2016",
    fill="country"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
#Plot 2
df2 = df %>%
  filter(is_canceled==0) %>%
  mutate(guests = adults+children+babies) %>%
  rowwise() %>%
  mutate(nights_stayed=stays_in_week_nights+stays_in_weekend_nights) %>%
  select(hotel, arrival_date, nights_stayed) %>%
  # Filter out only stayed for day, not night
  filter(nights_stayed>0) %>%
  mutate(start_date=arrival_date, end_date=arrival_date+nights_stayed-1)
dates=seq(min(df2$start_date), max(df2$end_date), by="days")
values = rep(0, length(dates))
names(values) <- dates
values2 = values
for (i in 1:nrow(df2)) {
  start = df2$start_date[i]
```

```
  end = df2$end_date[i]
  days = seq(start, end, by="days")
  for (j in seq_along(days)) {
    day = as.character(days[j])
    if (df2$hotel[i]=="Resort Hotel") {
      values[day] = values[day] + 1
    } else values2[day] = values2[day] + 1
  }
}
temp = data.frame(hotel="Resort Hotel", date=ymd(names(values)), count=values)
temp2 = data.frame(hotel="City Hotel", date=ymd(names(values2)), count=values2)
df3 = bind_rows(temp, temp2) %>%
  filter(year(date)==2016) %>%
  mutate(month = month(date, label=TRUE))
```

```
middle = seq(ymd("2016-01-15"), ymd("2016-12-15"), by="months")
max_city = max(df3$count)
df3.1 = df3 %>%
  group_by(hotel) %>%
  summarise(max_rooms = max(count)) %>%
  mutate(date=ymd("2016-01-01"))

ggplot(df3, aes(x=date, y=count)) +
  geom_line(mapping=aes(colour=hotel)) +
  geom_hline(data=df3.1, aes(yintercept = max_rooms, colour=hotel), linetype="dashed") +
  scale_x_continuous(breaks=middle, labels=month(middle, label=TRUE)) +
  geom_text(data=df3.1, aes(x=date, y=max_rooms,label = max_rooms, vjust = 1.2), size=3) +
  labs(
    x="Month",
    y="No. rooms occupied",
    title="Number of rooms occupied by month for 2016"
  )+
  theme(plot.title = element_text(hjust = 0.5))
```



Discussions

Plot1 From the bar graph, we can see the proportion of guests from each country varies from month to month. For example, the number of Spain guests is high in March, August and December, the number of guests from Britain is high in May, September and October. This could be due to the different cultures and practices in the various countries. We also see that the number of guests remains high from March to October, and low from November to February. This coincides with the Winter season in Portugal, indicating guests may prefer to come to Portugal during seasons other than Winter.

Plot2 From the visualization, we can see that the proportion of rooms occupied in both hotels is close to 100% for many months in the year, except for November to February, with the exception of late Dec to early January, likely because of Christmas. This follows closely with the trend of total guests per month, indicating room availability does affect the number of guests in the hotel. Interestingly, the number of guests staying in the hotel still varies quite significantly from month to month even when hotel capacity is almost full. One possible reason is that the hotels vary the price of the larger rooms to match demand. During peak season, large rooms will be occupied by more people. During off season, hotels might lower the price, letting smaller groups occupy large rooms to fill capacity.

Question 2: How does cancellation rates vary across different variable?

Introduction

The hotel cancellation rate is an important factor that affects the revenue of the hotel. Therefore, we wish to dive deeper into the dataset to uncover some of the causes and sources of these hotel cancellations. We aim to aid the business supervisor in making an informed decision when trying to reduce the hotel cancellation rates. In this question, we study how cancellation rates vary according to travel agents(Plot1) and the days on the waiting list(Plot2).

Methodology

Plot1 In this part, we study how cancellation rates vary for different travel agents. We first group the data according to the travel agent's ID and then get the cancellation rate, which is calculated by taking the number of cancelled bookings over the total number of bookings made by that agent. With this information, we then filter the agents with cancellation rates above 80% to be used in our plot. We decided to go for a scatterplot with the y-axis being the cancellation rate and the x-axis being the agent IDs. The individual points on the scatterplot are also labelled with the agent IDs to make it easier to identify the agent IDs. The colour gradient is also used on the points, where a darker colour signifies a higher cancellation rate. We have chosen a scatterplot as it gives a brief and clear overview of the agents who have high cancellation rates, while still maintaining the key information.

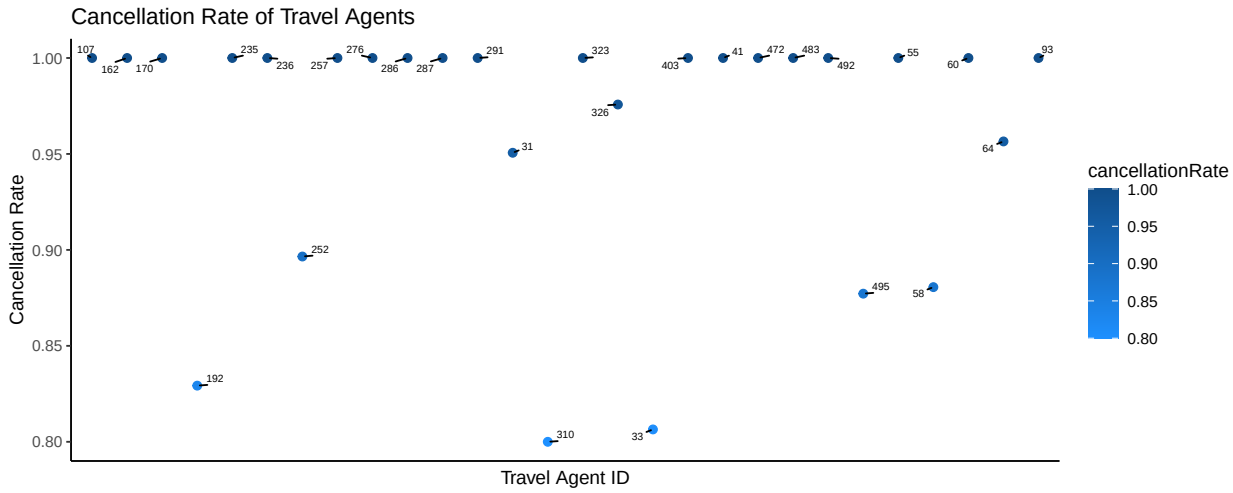
Plot2 In this visualization, we plotted a dodge bar to compare the cancellation rate when there were long waiting times versus short waiting times. Our group had defined more than 20 days on the waiting list as a long waiting time. Firstly, the data is mutated such that days on the waiting list are separated into long and short waiting times. The data is then grouped by the waiting times and filtered into whether the booking was cancelled or not. Next, the dodge bar is plotted such that the proportion of cancellations in both waiting times can be seen. Inclusive colours and a legend are also used such that red shows cancelled bookings and green shows bookings that have been successful.

Visualizations

```
# Question 2 Plot 1

# find agents with cancellation rates >= 80%
num_of_cancellations = df %>%
                        group_by(agent) %>%
                        filter(is_canceled==1) %>%
                        count()

num_of_bookings = df %>%
                    group_by(agent) %>%
                    count()

df_temp = num_of_bookings %>% inner_join(num_of_cancellations, by='agent')
df_temp = df_temp %>%
  mutate(cancellationRate = n.y/n.x) %>%
  arrange(desc(cancellationRate))
df_temp = filter(df_temp, cancellationRate >= 0.8)
#df_temp
```

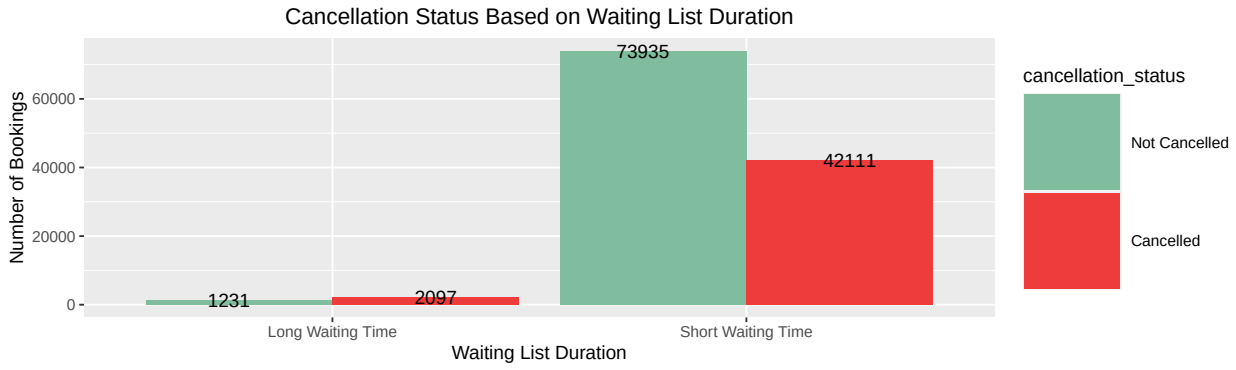
```
ggplot(df_temp, aes(y=cancellationRate, x=agent, label=agent)) +
  geom_point(aes(colour=cancellationRate), size=2) +
  scale_colour_gradient(low = "dodgerblue", high = "dodgerblue4") +
  geom_text_repel(min.segment.length = 0, size=2, max.overlaps = Inf) +
  labs(x= "Travel Agent ID",
       y= "Cancellation Rate",
       title = "Cancellation Rate of Travel Agents") +
  theme_classic() +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

Plot1



```r
# 20 days of waiting list is considered Long Waiting List, otherwise it's considered short.

waiting_list_hotels<-transform(
  df,
  waiting_list_time=ifelse(
    days_in_waiting_list > 20,
    "Long Waiting Time",
    "Short Waiting Time"
  )
)

waiting_list_distribution = waiting_list_hotels %>%
                              group_by(waiting_list_time, is_canceled) %>%
                              count()

waiting_list_distribution$cancellation_status= c("No", "Yes", "No", "Yes")

dodge_bar_waiting_list <- ggplot(
  waiting_list_distribution,
  aes(x = waiting_list_time, y = n))+
  geom_col(aes(fill = cancellation_status), position=position_dodge())+
  labs(title = "Cancellation Status Based on Waiting List Duration",
  x = "Waiting List Duration", y = "Number of Bookings")+
  theme(plot.title = element_text(hjust = 0.5),
        legend.key.size = unit(2, 'cm'),
        legend.key.height = unit(2, 'cm'),
        legend.key.width = unit(2, 'cm'))+
  scale_fill_manual(
    values = c("#80BD9E","brown2"),
    labels =c("Not Cancelled","Cancelled"))+
    geom_text(aes(label=n, group = cancellation_status),
              position = position_dodge(width = 1)
    )

dodge_bar_waiting_list
```

Plot2



Discussions

Plot1 The plot shows all the agents with cancellation rates greater than 80%. The higher the agent is in the plot, the higher its cancellation rate. We hope that the reader can identify quickly and easily the travel agents with high cancellation rates, and therefore be more careful when partnering with them. This plot serves as a reminder for the hotel business supervisor and highlights to him that there are indeed agents with extremely high cancellation rates, with some even hitting a 100% cancellation rate. Instead of blindly working with several travel agents to boost the number of bookings, the supervisor can make a better decision by choosing travel agents with good records, and one of the metrics that could be used in this case would be its cancellation rates.

Plot2 As seen from the dodge bar, there are a greater proportion of cancellations when the waiting time is longer. Long waiting times resulted in 63.0% of cancellations while short waiting times had 36.2% of cancellations. We hope that this information can clearly show the viewer that longer waiting times result in more cancellations. Given long waiting times, customers would likely look to other options and the hotelier would lose revenue. Moving forward, further investigation could be done to find out how long customers are likely to wait on a waiting list before cancelling their reservations. This information is likely to help the hotelier optimize how much resources and manpower to prepare in light of an influx of customers from the waiting list.

Reference

Data Source: Our data source is from The TidyTuesday Project. Link to the dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md

Seasonality in tourism demand. Seasonality in tourism demand - Statistics Explained. (n.d.). Retrieved April 17, 2022, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Seasonality_in_tourism_demand

Buchholz, K., & Richter, F. (2019, April 3). Infographic: Korea's Cherry Blossom Tourism Bump. Statista Infographics. Retrieved April 17, 2022, from https://www.statista.com/chart/17588/international-visitors-to-south-korea/